Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 138 (2018) 191–196

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2018

Establishment of Access Levels for Health Sensitive Data Exchange through Semantic Web

Vinicius Costa Lima[a,*], Domingos Alves[b], Felipe Carvalho Pellison[a], Vinicius Tohoru Yoshiura[c], Nathalia Yukie Crepaldi[b], Rui Pedro Chartes Lopes Rijo[d]

[a] *Bioengineering Postgraduate Program, University of São Paulo, São Carlos, Brazil*
[b] *Department of Social Medicine, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*
[c] *Community Health Postgraduate Program, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil*
[d] *School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal*

Abstract

Data exchange in health information systems must be carefully planned and needs to be protected from unauthorized access due to sensibility of stored content. Security aspects like authentication, authorization and encryption must be considered in this context. The main goal of this article is to present the implementation of security mechanisms to a semantic API that allows data extraction from a regional health information system designed to create notifications and to follow patients diagnosed with Tuberculosis. Data semantically tagged will be mapped individually to several access levels. It will be showed how external systems can connect, authenticate and retrieve only authorized data that are classified in the scope of its maximum access level.

Selection and peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: semantic web; health information system; health data; interoperability; security; API;

* Corresponding author.
E-mail address: viniciuslima@usp.br

Selection and peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.
10.1016/j.procs.2018.10.027

1. Introduction

The monitoring of clinical data and notification of cases related to Tuberculosis (TB), a bacterial infectious disease that mainly affects the lungs and still remains as one of the major global health problems [1], is an example that represents the importance of using information systems in health area. In Brazil, TB cases notification are compulsory [1] [2]. The patients' follow-up requires filling several standardized and specific forms usually according to Directly Observed Treatment Short Course (DOTS) strategy. The data collected must be inserted, in some cases, in more than one system, resulting in unnecessary data duplication and difficulties to offer integrated treatment in different health levels. It is essential to integrate several sources of available data to improve data quality, so that it can be transformed in knowledge through reducing incompleteness, inconsistency and duplicity.

To achieve better support on decision-making processes, both on operational and administrative levels, health systems must be able to exchange data. In other words, systems must support interoperability, formally defined as the capacity of one or more systems exchange data and use it in a transparent way [4]. To do that, there are standards and protocols that should be used as references. In addition, this concept can be divided, mainly, in two subtypes, such as functional and semantic interoperability. Functional interoperability is related to a group of rules that must be applied during the communication process and information exchange between systems. Semantic interoperability, in turn, is related to preserving the semantic value of a given data to allow information to be correctly understood by another system or application that was not initially developed for the same purpose [5] using vocabularies defined to specific domains. The combination of these two types of interoperability can add an essential feature to health information systems, increase data integration and reduce the duplicity and data heterogeneity.

In this scenario, the World Wide Web Consortium (W3C) plays a significant role by defining standards to be used on the Semantic Web. The Semantic Web can be defined as an extension of current web aimed to provide a better information treatment by adding semantic value to data. Through the Semantic Web, it is possible to link content to a specification provided by ontologies, which can be defined as a formal representation of knowledge in a specific domain, thus, formulating a rigorous and exhaustive conceptual scheme [6] [7].

However, data exchange in health information systems must be carefully planned and needs to be protected from unauthorized accesses. Sensitive content that involve identity of individuals or that could harm the rights of privacy or honor are classified as confidential. Therefore, mechanisms such as data access authorization levels and safe data transfer should be considered to enhance security in the context of sensitive data exchange between systems.

The main goal of this article is to present the implementation of security mechanisms to an API (Application Programming Interface) that allows data extraction from a regional health information system used to register and follow patients diagnosed with Tuberculosis. The API is prepared to semantic interoperability, i.e., is able to send responses with semantic value added in standardized semantic formats. It will be showed how external systems can connect, authenticate and retrieve only authorized data that are classified in the scope of the maximum access level assigned to the external system.

In the next section, the concepts and environments that will be used as references to support the development of this work will be presented. In the third section, the methodological approach, along with the selected technological tools will be described. In the fourth section, the results achieved by using the background knowledge and by applying the proposed methodology will be showed. Finally, expected contributions will be highlighted in the fifth section, as well as future work.

2. Background

2.1 Security on Semantic Web

The Semantic Web can be considered an evolution or an extension of the current web [8]. Its purpose is to add semantic value to contents on the web. This is achieved by tagging the content with semantic markups making them easily accessible by other systems and still remain interpretable by humans through the classical representation of the current web. Semantic Web allows data interpretation by machines without worrying about its form of representation [7].

There are some pillars that make up the Semantic Web paradigm: a data model; a query protocol; and a set of reference vocabularies. Standards preconized by W3C such as Resource Description Framework (RDF,

https://www.w3.org/RDF/), SPARQL Protocol and RDF Query Language (https://www.w3.org/TR/rdf-sparql-query) and ontologies are related to these pillars and can provide the ability to add and retrieve semantic value to/from a dataset. RDF can be defined as a description language and a model for representing and exchanging semantic data on the web. SPARQL is defined as a language to perform queries over data represented as RDFs. Ontologies are defined as a formal specification of concepts in a domain of interest [6] [9] used to tag contents on the web that can be represented by the properties defined in an ontology.

Other protocols, tools and abstract layers also compose the Semantic Web. In the scope of this article, URIs (Uniform Resource Identifiers), Microdata (https://www.w3.org/TR/microdata/) and layers related to security aspects were also considered. Ontologies and entities semantically tagged are identified by URIs. Microdata is a technology that extends HTML attributes to allow semantic markups in web pages. Finally, layers such as Trust and Cryptography can provide the necessary support and guideline to implement security mechanisms.

Patients related data (personal information, treatment records, etc.) must remain confidential and can be shared only with authorized systems that meet security requirements. Semantic Web does not provide such implemented security algorithms, so it must be deployed to the system environment as additional features. Thus, recognizing the importance of developing standards and tools to enhance security on Semantic Web [10], the establishment of data access levels, the main goal of this paper, and the protection of confidential data can be included in security related layers [10].

In this work, sensitive data were protected by exchanging them under an encrypted channel and by developing an API with authentication and authorization mechanisms to provide access to identified systems and to segment data in different levels, so that each external system can reach only information that are in the scope of its assigned maximum access level.

2.2 SisTB

SisTB is a regional web-based system for control of TB patients. Its main functionalities are patients registration and treatment follow up through DOTS strategy.

This system is part of a computational health ecosystem that consists on a set of integrated systems that aims to do better management and exchange of information related to TB cases [11]. Additionally, there is a functional and semantic interoperability architecture that contemplates SisTB, making it able to exchange relevant data with authorized systems [11]. The HTML pages from SisTB are semantically tagged, allowing data to be extracted on the fly with content being dynamically generated. Thus, due to these features, SisTB will be the main data source for this project.

3. Research Methods

The scientific research methodology basis for this work is the Action Research. It is a suitable methodology because it can be used to deal with quick issues or as an intelligent procedure for dynamic critical thinking through the main procedure's motivation, that is, to address a specific issue and to create rules for best practice [12]. This strategy is supported since the project has a practical component, besides the theoretical research development.

3.1 Overview

The novelty of this paper is the establishment of access levels for health data exchange along with a semantic API based on the Semantic Web paradigm. Through the access levels, it will be possible to segment semantically tagged content on SisTB to allow access only to specific systems that have an adequate access level. Systems that do not meet this access level will not be authorized to view the information. Additionally, external systems will be identified and authenticated through public/private key pairs.

The first step was the evaluation on how semantic interoperability could be achieved through Semantic Web considering the desired security aspects and existent architecture [11]. Next, the strategy to implement the authentication mechanism (external system identity validation) was defined, along with the access levels and how them would be assigned to existing data. Finally, the most suitable technological tools were selected.

The use of HTTPS protocol is also essential to provide confidentiality over the communication channel that has been used to exchange sensitive semantic data through the web. Hence, if any messages got intercepted between the requester and the API, the content will remain protected due to the encryption implemented by the mentioned protocol.

The authentication mechanism will be implemented using a public/private key pair. External systems must send in their requests to the API a signature generated with a combination of parameters and its private key. When the request is received, the API will check the signature consistence with the public key previously installed on a repository key.

As part of the strategy, levels were assigned to ontology properties and not directly to the data. By doing that, we can easily apply the access level to all semantically tagged data with a given ontology. The mapping of properties to access level will be stored in a relational database so it can be queried to verify each property obtained from a tagged HTML page. Four access levels were defined and are represented by colors. From lower to higher level, the colors are respectively: green, yellow, red and black. The green is the lower level and is the default value for all properties. Yellow is the intermediate level while red represents the higher level. Properties assigned with black level will be always censured and will not be transferred to the requester. Fig. 1 illustrates these access levels.

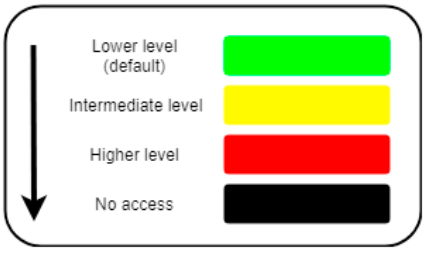

Fig. 1 - Available Access Levels

3.2 Technological tools

This project was developed using tools that were chosen considering mainly its maturity and compatibility with Semantic Web resources and with Java programming language. Furthermore, all selected tools are open-source.

The semantic API, including the security layer, was developed using Java, version 8, since there are semantic frameworks that helps the creation of semantic applications like Apache Jena (https://jena.apache.org/) and Apache Any23 (https://any23.apache.org/) libraries. The first one provides all interfaces, classes and methods necessary to work with semantic data in RDF and others semantic formats [13]. The second one is used to extract data from semantically tagged HTML pages and to transform them into RDF format.

In order to tag HTML pages and to add semantic value to contents, ontologies are strictly necessary. Although SisTB makes use of a set of ontology, the ontology Person defined by schema.org community (http://schema.org/Person) was selected to develop a proof of concept regarding the main goal of this project. This is due to its generic and wide scope of application and its simplicity. In the case of SisTB, patients' personal data were tagged using the Person ontology. Microdata was the chosen tool to perform semantic markups on SisTB HTML pages.

Finally, MySQL (https://www.mysql.com/) relational database was selected to store security metadata (called security repository) to external systems authentication and authorization.

4. Results

Through the development of an API to SisTB, it is possible to send responses in semantic formats (RDF and JSON-LD) and implement security aspects, enabling not only the data extraction from semantically tagged HTML pages by authorized systems, which are segmented in distinct access levels, but also allowing functional and semantic interoperability to health data related to TB cases. Requesters do not need to know which HTML page contains a specific data because the API is responsible to mapping URL endpoints to corresponding web pages.

SisTB already has semantic tagging in HTML web pages in sections like patients' personal data, treatment information. administrative indicators and others. For the intended proof of concept, the patient's personal records section was tagged using Person Ontology from schema.org community.

The data extraction is performed through HTTPS requests using GET method to a desired API endpoint. For instance, the request should be made to an URL like:

https://my.domain/sistb-api/semantic/patient/<id>?format=<rdf|json-ld>&signature=<signature>

The value *id* is a unique identifier for a patient, while format parameter represents the expected format to be received in the response (RDF or JSON-LD) and *signature* parameter is related to the authentication mechanism. Fig. 2 shows an overview of the communication flow between the proposed API and external systems.

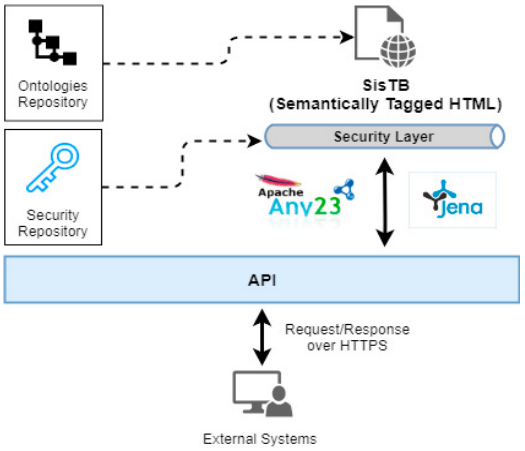

Fig. 2 - Communication flow overview

Each external system must have a pair of keys (public and private). They need to include in their requests a signature build with a combination of parameters and its private key for the purpose of authentication (identity validation). Despite of not being the preferred mechanism, HTTP basic authorization method, i.e. the inclusion of a pre-shared username and password in the request header, is supported, since the communication channel is over HTTPS and the credentials will remain safe.

Once the API receives the request, the security layer will start checking the credentials. If a signature was provided, it will be validated through the external system public key previously installed on a security repository. In case of authentication fails, a HTTP response with 401 code (Unauthorized) is sent. Apart from that, the request is processed and by API.

Using Apache Any23 library, the API extracts the requested data from a HTML page and builds a semantic model based on RDF which will incorporate the content (represented as properties) semantically tagged and the metadata that define its meaning. After that, the model will be analyzed by the security layer through Apache Jena library, when each property will be checked individually to verify if the requester is authorized to retrieve them.

If there is an entry in the security repository related to a property, the access level assigned will be confronted with the access level allocated to the external system. Otherwise, it will be considered as level *green*, that represents the lower access level to which all external systems have access. In case of the system has the equivalent or higher access level, the property and its value will be included in the model. On the other hand, the property will be censured. In addition, if a property has a *black* access level assigned, it will be censured immediately apart of the external system maximum access level. When a property is censured, it will be still included in the model, but its value will be hidden.

After verifying all properties carefully, the response is sent to the requester over the secure communication channel in the format informed in the request URL. Fig. 3(a) shows an example of response in JSON-LD format with censorship for an external system with maximum access level defined as *green* (lower). Fig. 3(b) shows an example of response in JSON-LD format without censorship for an external system with access level defined as *red* (higher). Properties *gender*, *birthDate*, *birthPlace* have *green* level assigned. *Yellow* level was allocated to *telephone* and *address* properties. Finally, *givenName* property is assigned with *red* level.

(a)
```
{
  "@graph" : [{
    "@id" : "_:b0",
    "@type" : "http://schema.org/Person",
    "address" : "INSUFFICIENT ACCESS LEVEL",
    "birthDate" : "30/04/2000",
    "birthPlace" : "RIBEIRAO PRETO",
    "gender" : "Masculino",
    "givenName" : "INSUFFICIENT ACCESS LEVEL",
    "telephone" : "INSUFFICIENT ACCESS LEVEL"
  }],
  "@context" : {
    //omitted content
  }
}
```

(b)
```
{
  "@graph" : [{
    "@id" : "_:b0",
    "@type" : "http://schema.org/Person",
    "address" : "Rua 105 - RIBEIRAO PRETO - SP",
    "birthDate" : "30/04/2000",
    "birthPlace" : "RIBEIRAO PRETO",
    "gender" : "Masculino",
    "givenName" : "NOME TESTE 10",
    "telephone" : "(16) 1051-0510"
  }],
  "@context" : {
    //omitted content
  }
}
```

Fig. 3 - Semantic response in JSON-LD format with (a) censorship (b) full content

5. Final Considerations

Semantic Web allied to classical technological resources like APIs can empower systems with semantic and functional interoperability with low cost adaptations. However, regarding the use and exchange of health sensitive data, additional features must be implemented to ensure security during the communication process.

Therefore, the establishment of data access levels in semantic interoperability is the expected result of this article. By doing that, besides of security aspects like authenticating external systems to validate identity and the use of HTTPS protocol to provide confidentiality, is possible to segment data and share it only with entities with a granted maximum access level. Furthermore, with the purpose of easily assigning access levels to data (also called properties), all security metadata is stored in a repository. Data can be retrieved from SisTB through semantically tagged HTML pages. The API is able to send responses in semantic formats, such as RDF and JSON-LD. All data that are not allowed to be shared will be censured.

As future work, the configuration of a SPARQL endpoint is desired because it can increase possibilities of interaction with the API, facilitating information attaining regardless of marked up web pages. The development of an administrative portal to manage security metadata information is also expected.

Acknowledgements

This work was developed within the Bioengineering Postgraduate Program (Programa de Pós-Graduação Interunidades em Bioengenharia - EESC/FMRP/IQSC) of University of São Paulo.

References

[1] M. da S. BRASIL, “Doenças infecciosas e parasitárias: guia de bolso,” *Cad. Saude Publica*, vol. 22, no. 11, pp. 2498–2498, 2010.
[2] M. da S. BRASIL, *Portaria nº 204*. 2016.
[3] L. D. de Sá *et al.*, “Implantação da estratégia DOTS no controle da Tuberculose na Paraíba: entre o compromisso político e o envolvimento das equipes do programa saúde da família (1999-2004),” *Cien. Saude Colet.*, vol. 16, no. 9, pp. 3917–3924, 2011.
[4] A. Geraci, F. Katki, L. McMonegal, B. Meyer, and H. Porteous, *IEEE Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries*. 1991.
[5] European Comission, “COMMISSION RECOMMENDATION of 2 July 2008 on cross-border interoperability of electronic health record systems,” 2008.
[6] H. Liyanage, P. Krause, and S. de Lusignan, “Using ontologies to improve semantic interoperability in health data,” *J. Innov. Heal. Informatics*, vol. 22, no. 2, pp. 309–315, 2015.
[7] I. Robu, V. Robu, and B. Thirion, “An introduction to the Semantic Web for health sciences librarians.,” *J. Med. Libr. Assoc.*, vol. 94, no. 2, pp. 198–205, 2006.
[8] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities,” *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.
[9] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.
[10] B. Thuraisingham, “Security standards for the semantic web,” *Comput. Stand. Interfaces*, vol. 27, no. 3, pp. 257–268, 2005.
[11] F. C. Pellison *et al.*, “Development and evaluation of an interoperable system based on the semantic web to enhance the management of patients’ tuberculosis data,” *Procedia Comput. Sci.*, vol. 121, pp. 791–796, 2017.
[12] D. J. Greenwood and M. Levin, “Introductions to Action Research: Social Research for Social Change,” p. 315, 2007.
[13] M. Lanthaler and C. Gütl, “On using JSON-LD to create evolvable RESTful services,” *Proc. Third Int. Work. RESTful Des. - WS-REST ’12*, p. 25, 2012.